

A Block-Based Triangle Counting Algorithm on Heterogeneous Environments

Abdurrahman Yaşar, Sivasankaran Rajamanickam (*Member, IEEE*), Jonathan Berry, and Ümit V. Çatalyürek (*Fellow, IEEE*)

Abstract—Triangle counting is a fundamental building block in graph algorithms. In this paper, we propose a block-based triangle counting algorithm to reduce data movement during both sequential and parallel execution. Our block-based formulation makes the algorithm naturally suitable for heterogeneous architectures. The problem of partitioning the adjacency matrix of a graph is well-studied. Our task decomposition goes one step further: it partitions the set of triangles in the graph. By streaming these small tasks to compute resources, we can solve problems that do not fit on a device. We demonstrate the effectiveness of our approach by providing an implementation on a compute node with multiple sockets, cores and GPUs. The current state-of-the-art in triangle enumeration processes the Friendster graph in 2.1 seconds, not including data copy time between CPU and GPU. Using that metric, our approach is 20 percent faster. When copy times are included, our algorithm takes 3.2 seconds. This is 5.6 times faster than the fastest published CPU-only time.

Index Terms—triangle counting, task-based, block-based, sub-graph, multi-core, multi-GPU.

1 INTRODUCTION

Graphs are very useful data-structures. They can represent different applications, such as social network analytics, biological networks, and scientific simulations. Today there exist large graphs that have billions of vertices and edges. High-performance processing of these large graphs is crucial and pervasive. Most of the current high-performance solutions rely on distributed systems because one can process problem instances that are bigger than the main memory of a single node. However, communication and data replication in distributed systems often leads to bottlenecks. The popularity of multi-core machines increased drastically during the past decade. Today, different hardware vendors have developed processors with multiple cores to deliver improved performance, and hence the multi-core technology became ubiquitous. In addition to multi-core technology, many hardware accelerators like graphics processing units (GPUs), and field-programmable gate arrays (FPGAs) also became a part of the systems to serve different parallelization needs, and more are coming. Soon, we will have distributed systems consisting of multi-core servers with many different types of hardware accelerators. Such an environment increases the importance of designing flexible algorithms for performance-critical kernels and implementations that can run well on various platforms. In this work, we propose an architecture-agnostic triangle counting algorithm, BBTC, and we do an extensive analysis

of its performance on a single node multi-CPU and multi-GPU execution environment.

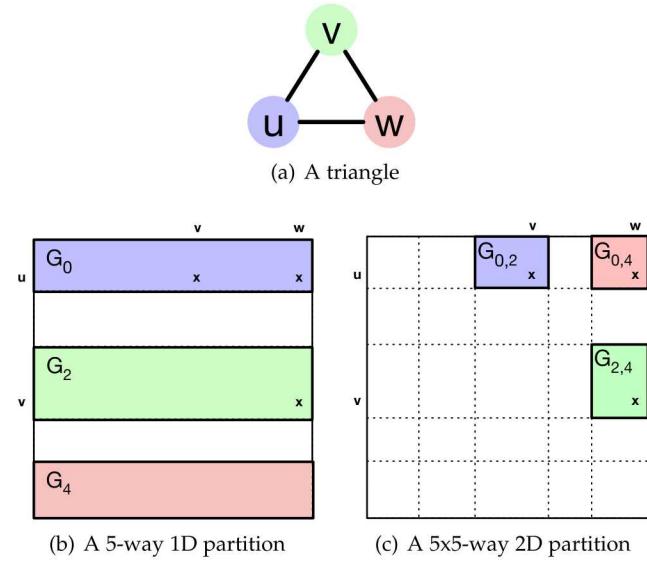


Fig. 1. Finding graph triangles in 1D and 2D-partitioned adjacency matrices.

The triangle counting problem [1]–[9] is to find the number of 3-cliques in an undirected graph (Fig. 1(a)). This is a crucial graph kernel that serves as a building block for many other graph problems such as k-truss decomposition [10], community detection [11], thematic structure identification [12], dense sub-graph discovery [13] and link recommendation [14]. In recent years, algorithmic and architectural advances have led to great improvements in triangle counting performance [3], [15]–[17]. None of these approaches are truly heterogeneous because they all target one architectural feature. In Section 4, we propose a block-

• Yaşar, and Çatalyürek are with the School of Computational Science and Engineering at Georgia Institute of Technology.

E-mail: {ayasar,umit}@gatech.edu

• Rajamanickam, and Berry are with the Center for Computing Research at Sandia National Laboratories, Albuquerque, NM.

E-mail: {srajama,jberry}@sandia.gov

Manuscript received xx xx, xx; revised xx xx, xx.

based approach for heterogeneous execution environments that simultaneously leverages both CPUs and multiple GPUs. Such an approach is crucial for achieving three goals: maximizing memory utilization, processing graphs that cannot fit into GPU memory and overlapping the data-copy time with the computation. To achieve these goals, we tackle the triangle counting problem under three axes; data/computation partitioning, a block-based formulation for the triangle counting problem, and architecture agnostic task execution.

At the abstract level, our execution model is as follows. Computation is divided into tasks. Each task depends on multiple blocks. A light-weight scheduler schedules tasks on available devices. Scheduler is also responsible for the movement of data blocks that are needed for execution of each individual task.

Execution time is a function of the computational load of individual tasks, and communication costs. The granularity of the task decomposition is crucial for performance. Therefore, partitioning is the first step and refers to both data and computation partitioning. In the context of graphs, we can do the partitioning at the vertex level (coarse-grained) or the edge level (fine-grained). A hybrid approach may also be used. There is a duality between graphs and sparse matrices. Vertex-level and edge-level partitionings correspond to row (1D) and nonzero (2D) partitionings. For visualization purposes, consider the adjacency matrix representation of a graph. Fig. 1(b) illustrates a 1D partitioning. This generates coarse-grained tasks. Some state-of-the-art CPU-based triangle counting algorithms (such as [2], [18]) generate that kind of coarse-grained tasks. However, such a task generation ends up with higher copy/communication costs on a heterogeneous environment and highly imbalanced workload distribution [19]–[23]. We can address this problem by using a 2D partitioning strategy and generating medium-grained tasks (e.g., see Fig. 1(c)) to solve the triangle counting problem. Generally, there are two classes of algorithms that provide such partitioning: i) connectivity-based techniques (graph and hypergraph) [21], ii) spatial partitioning techniques [20]. Connectivity-based techniques find a reordering of the vertices while spatial techniques break nonzeros of the matrix into blocks without reordering the vertices.

We seek partitioning methods that maximize memory efficiency and minimize data movement which is essential to overlap the communication with the computation. Unfortunately, unstructured or random partitionings may easily end up with highly irregular data access patterns (i.e., one block may require to access all other blocks). Our ultimate goal is to localize sub-graph computations, by improving data access patterns and data movement in the system architecture. Therefore we need a special partitioning scheme that we could use for divide-and-conquer approaches. For this reason, we utilize a special 2D partitioning scheme called *symmetric rectilinear partitioning* (also known as *symmetric generalized block distribution* (see Fig. 1(b)). This problem, first defined by Grigni and Manne [24]. Recently, Yaşar and Çatalyürek [25] proposed some heuristic algorithms to solve this problem. It is also possible to achieve such partitioning by utilizing more complex connectivity-based [21] techniques. However, these techniques have two drawbacks;

first, most of these techniques are computationally expensive, second, connectivity-based ordering of the vertices may affect the computational complexity of the algorithm. In sparse graphs with heavy-tailed degree distributions, triangle counting has low computational complexity [5] therefore we need a fast partitioning technique. In this work, we use a parallel version of the PBD algorithm presented in [25] to partition the graph into blocks.

We propose a medium-grained triangle counting formulation (BBTC) on top of symmetric rectilinear partitioning. This makes the algorithm naturally suitable for task-based execution on shared and distributed-memory systems as well as on heterogeneous architectures. BBTC defines tasks based on generated blocks. CPUs and GPUs process tasks together. Although the creation of tasks is dependent on graph data, our algorithm to dispatch and execute these tasks is data-agnostic. This agnostic structure will allow execution on heterogeneous settings with any accelerator and CPU combination.

One can further optimize the execution time by utilizing different data structures and implementation of the graph kernel based on properties of the input data and architectures. However, in this work, we utilize a single data-structure and just one implementation of each task type on each processor type. The power of the algorithm comes from partitioning of the triangle space into tasks, a dynamic task-scheduling scheme that schedules tasks on multi-core-CPUs, use of multiple streams on multiple GPUs to effectively utilize the massive computing capabilities on the GPUs, and, at the same time, move the data to the GPUs asynchronously.

The primary contributions of this work can be listed as:

- We show how to partition a graph's triangle set, and we use this observation to derive a new block-based triangle counting algorithm.
- We show how to run our algorithm on heterogeneous hardware when the graph does not fit on device.
- We propose dynamic scheduling techniques for block-based execution model for heterogeneous environments and show how tasks can be distributed among different CPUs and GPUs efficiently.
- We evaluate characteristics of our algorithm in theoretical and practical aspects and perform extensive analysis on different architectures.

The preliminary results of this work were presented as an extended abstract in [16]. This paper presents a detailed description of our algorithm and its complexity analysis, a thorough experimental evaluation of the components of the algorithm and comparisons against the state-of-the-art. These results demonstrate that our implementation beats the fastest published end-to-end CPU execution times. The performance improvement is up to 11 \times over our previous state-of-the-art implementation [18], 16 \times over another state-of-the-art implementation [2], and 1.5 \times over state-of-the-art multi-GPU implementation [3] on the largest three graphs.

2 PROBLEM FORMULATION

An undirected graph $G = (V, E)$, consists of a set of vertices V and a set of edges E . An edge e is referred

to as $e = (u, v) \in E$, where $u, v \in V$. In the context of triangle counting, efficient algorithms traverse vertices and edges following a complete ordering, which avoids double counting and also reduces the number of edges traversed. For the sake of simplicity, in this work, we will use vertex IDs for that complete ordering in the traversal. Since G is undirected, we represent each edge as (u, v) , where $u < v$. Under this representation, we call u the *source* vertex and v the *destination* vertex. Thus, we induce an ordering that is helpful when explaining our partitioning. The neighbor list of a vertex $u \in V$ is also defined using the same complete ordering, i.e., $N(G, u) = \{v \in V : (u, v) \in E \text{ and } u < v\}$. The degree of a vertex $u \in V$ is defined as $d(G, u) = |N(G, u)|$. We will use n and m for number of vertices and edges, respectively, i.e., $n = |V|$ and $m = |E|$.

Definition 1. Triangle Counting Problem. Given an undirected graph $G = (V, E)$, the triangle counting problem computes the number of unique three cliques, i.e., the number of mutually connected three vertices, in the graph, G .

Given a graph G , and an integer p such that $1 \leq p \leq n$. Let V^p be a vertex partition set of V into p non-overlapping partitions such that; $V^p = \{V_0, \dots, V_{p-1}\}$ where $\cup_{0 \leq i \leq p-1} V_i = V$ and $\cap_{0 \leq i \leq p-1} V_i = \emptyset$.

Let $G_{i,j}$ be a sub-graph of G such that $G_{i,j} = \{(u, v) \in E : u \in V_i \wedge v \in V_j\}$. By definition $G = \cup_{0 \leq i, j \leq p-1} G_{i,j}$. Our $G_{i,j}$ construct allows us to avoid double-counting any triangles. In the partitioned graph we will only use sub-graphs ($G_{i,j}$) that appear in upper triangular part of the adjacency matrix, ($i \leq j$) which allow us to only traverse edges $(u, v) \in G_{i,j}$, where $u < v$ by definition. Source and destination vertex sets in a given sub-graph $G_{i,j}$, are defined as $V_s(G_{i,j}) = V_i$ and $V_d(G_{i,j}) = V_j$, respectively. The partial neighbor list of a vertex $u \in G_{i,j}$ is defined as $N(G_{i,j}, u)$. The partial degree of a vertex $u \in V_s(G_{i,j})$ is defined as $d(G_{i,j}, u) = |N(G_{i,j}, u)|$. All the notations are summarized in Table 1.

TABLE 1
Notation used in this paper.

Symbol	Description
$G = (V, E)$	A graph G with vertex and edge sets, V and E , respectively
$n = V $	Number of vertices
$m = E $	Number of edges
V^p	Vertex partition set; $V^p = \{V_1, \dots, V_p\}$
$V_s(G_{i,j}) = V_i$	Source vertices of the subgraph $G_{i,j}$
$V_d(G_{i,j}) = V_j$	Destination vertices of the subgraph $G_{i,j}$
$N(G, u)$	Neighbor list of vertex $u \in V$
$N(G_{i,j}, u)$	Partial neighbor list of vertex u , i.e., $N(G_{i,j}, u) = \{(u, v) : u \in V_i \wedge v \in V_j\}$
$d(G, u)$	Degree of vertex u , $d(G, u) = N(G, u) $
$d(G_{i,j}, u)$	Partial degree of vertex $u \in V_s(G_{i,j})$, $d(G_{i,j}, u) = N(G_{i,j}, u) $
t	A task; $t = \{i, j, k\} = \{G_{i,j}, G_{j,k}, G_{i,k}\}$ where $i \leq j \leq k$
$T = \{t_0, \dots, t_i\}$	Set of all tasks

Algorithm 1: NINTERSECT (A, B)

```

▷ Computes intersection count.  $A$  and  $B$  are sorted.
 $a \leftarrow 0; b \leftarrow 0; c \leftarrow 0;$ 
while  $a < |A|$  and  $b < |B|$  do
  if  $A[a] = B[b]$  then
     $\quad ++c; ++a; ++b;$ 
  else if  $A[a] < B[b]$  then
     $\quad ++a$ 
  else
     $\quad ++b$ 
return  $c$ 

```

Algorithm 2: TC-INTERSECT (G)

```

 $\tau \leftarrow 0$  ▷ Initialize number of triangles to 0
for each  $u \in V$  do
  for each  $v \in N(G, u)$  do
     $\quad \tau \leftarrow \tau + \text{NINTERSECT}(N(G, u), N(G, v))$ 
return  $\tau$ 

```

3 WORK OPTIMAL TRIANGLE COUNTING ALGORITHMS

Sequential triangle counting is a well-studied problem [1], [5]. The best algorithms rely on matrix multiplication and run in $O(n^\omega)$ or $O(m^{\omega/(1+\omega)})$ where ω is the matrix multiplication exponent [6], [8]. However, these algorithms are impractical and require $\theta(n^2)$ space. The fastest algorithm that doesn't require $\theta(n^2)$ space does $O(m^{3/2})$ work [1]. Merge-based and hashmap-based versions can both result in work optimal algorithms. Both algorithms process vertices based on their degrees to bound the maximum degree in the graph and guarantee $O(m^{3/2})$ work. This bound can be improved if the degree distribution is governed by a power law with exponent α . In such a case, Latapy [1] shows that triangle enumeration is bounded by $\Theta(mn^{\frac{1}{\alpha}})$. This is an asymptotic improvement over the worst-case. Berry, et al. [5] improved upon this further and show that complexity of this problem can actually be $\Theta(n)$ in realistic circumstances where the 4/3 moment of a graph is bounded by a constant.

Alg. 1 outputs the number of the common elements in given two sorted lists using the set intersection. This algorithm can be used in triangle counting if the neighborhood of each vertex is sorted based on the vertex IDs. While this algorithm is cache-friendly, it results in poor performance when there are high degree vertices. Alg. 2 outputs the number of triangles in the graph using this list-based intersection. Alg. 3 uses a hash map to count the number of common neighbors in two vertices' adjacency list.

Variations of Alg. 2 and Alg. 3 are widely used in many sequential and parallel works [1], [2], [18]. Latapy [1] proposed a refinement technique inspired from *ayz-listing* that doesn't require $\theta(n^2)$ space. This refinement technique applies a hash map based algorithm for high-degree vertices and a list-based algorithm for low-degree vertices. In this paper, we use Latapy's algorithm as our sequential baseline since it provides the fastest sequential execution.

Algorithm 3: TC-HMAP (G)

```

 $\tau \leftarrow 0$             $\triangleright$  Initialize number of triangles to 0
for each  $u \in V$  do
  for each  $v \in N(G, u)$  do
     $H[v] \leftarrow u$ 
  for each  $v \in N(G, u)$  do
    for each  $w \in N(G, v)$  do
      if  $H[w] = u$  then
         $\tau \leftarrow \tau + 1$ 
return  $\tau$ 

```

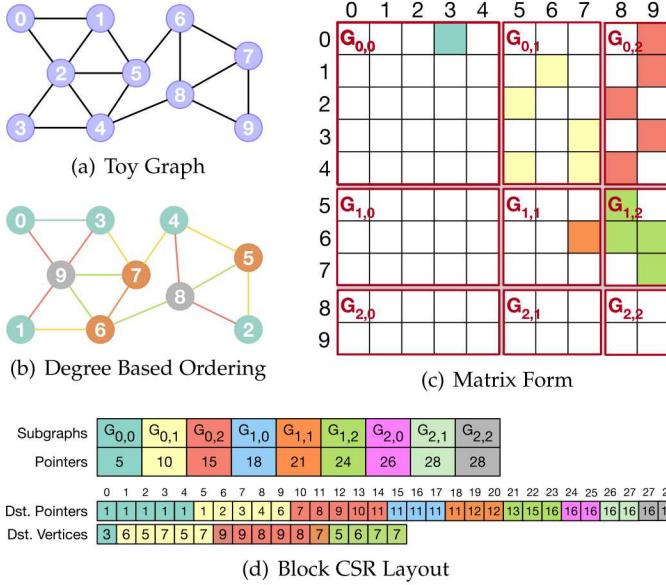


Fig. 2. Toy Example. 2(a): a toy graph. 2(b): Degree-based ordering of 2(a), each edge in a block colored with same color and each vertex in a vertex partition colored with the same color. 2(c): Adjacency matrix representation of the symmetric rectilinear partitioned graph. 2(d): Block CSR representation of the given partitioned graph. Tasks, $T = \{\{0, 0, 0\}, \{0, 0, 1\}, \{0, 0, 2\}, \{0, 1, 1\}, \{0, 1, 2\}, \{0, 2, 2\}, \{1, 1, 1\}, \{1, 1, 2\}, \{1, 2, 2\}, \{2, 2, 2\}\}$.

4 BLOCK-BASED TRIANGLE COUNTING (BBTC)

Most of the current state-of-the-art on the triangle counting problem require access to the entire adjacency list for each vertex. In this paper, we investigate the usage of partial adjacency lists (i.e., subgraphs).

4.1 Subgraph generation

A contiguous rectangle in an adjacency matrix corresponds to an *edge-induced subgraph* (a subset of edges in the graph). We propose to use symmetric rectilinear partitioning (also known as symmetric generalized block distribution [24]) to generate these subgraphs. This partitioning is similar to 2D matrix distributions that have been widely used in dense algebra with cartesian distributions [26] (i.e., the same partitioning vector is used for row and column partition) and also applied to sparse matrices [27]. In this partitioning, diagonal blocks are required to be squares.

As stated earlier, in order to avoid double counting and reduce the amount of edge traversal, we use a total ordering of vertex IDs and also only store upper triangular part of the adjacency matrix. Furthermore, in order to reduce

the number of edges traversed even further, similar to existing efficient sequential algorithms [1], [5], [7], [9], we re-order vertices in non-decreasing degree order, then apply symmetric partitioning to the adjacency matrix after these transformations. Figure 2 illustrates this process on a toy graph.

Again, by definition, edges of a triangle can appear in at most three of the subgraphs defined by our block-based partitioning. Furthermore, in a triangle, for a chosen pair of edges, there exists only one sub-graph such that the third edge belongs. Therefore we can bound the computational complexity of triangle counting and the data movement cost using symmetric rectilinear partitioning.

Our algorithm can work with any valid symmetric partitioning. The only constraint is the limited memory of the computing devices that will be used. Therefore, ideal symmetric partitioning should limit the sizes of subgraphs, such that three subgraphs can fit into memory of the computing devices. In this work, we used a lightweight symmetric rectilinear partitioning algorithm called PBD [25]. After the generation of the blocks, we use a block variant of CSR (BCSR) [28] to store the graph. This storage format is illustrated in Fig. 2(d). In this layout, blocks are ordered in row-major order.

4.2 Composition of the task list

Assume that vertices u, v and w form a triangle in a given graph (see Fig. 1). Let the first edge, (u, v) , appear in the subgraph $G_{i,j}$. It means that $u \in V_i$ and $v \in V_j$. Hence, the second edge, (v, w) , may appear in any subgraph $G_{j,k}$ where $j \leq k \leq p-1$ so $w \in V_k$. Therefore, by definition the third edge, (u, w) appears in $G_{i,k}$ because $u \in V_i$ and $w \in V_k$. In the context of this paper a task can be defined as below.

Definition 2. Task. Given a symmetric rectilinear partitioned graph, $G = (V, E)$ that is partitioned into $p \times p$ blocks, and a triplet $\{i, j, k\}$ such that $i \leq j \leq k$, a task, t , is defined as the unit of work to count the number of triangles in $\{G_{i,j}, G_{j,k}, G_{i,k}\}$.

To illustrate, in Fig. 2(b), vertices 3, 7 and 9 form a triangle in the toy graph. The first edge, $(3, 7)$, appears in the subgraph $G_{0,1}$, the second edge, $(7, 9)$, appears in the subgraph $G_{1,2}$ and the third edge, $(3, 9)$ appears in the subgraph $G_{0,2}$. This would be counted when $G_{0,1}, G_{1,2}$ and $G_{0,2}$ form a task; $t = \{0, 1, 2\} = \{G_{0,1}, G_{1,2}, G_{0,2}\}$.

One can count number of triangles, (u, v, w) , in given partitioned graph by considering all possible triples of blocks (i.e., tasks); $\{G_{i,j}, G_{j,k}, G_{i,k}\}$ where $(u, v) \in G_{i,j}$, $(v, w) \in G_{j,k}$ and $(u, w) \in G_{i,k}$. However in an undirected graph this approach counts each triangle six times. Most state-of-the-art work treats undirected graphs as directed by considering edges from smaller vertex IDs to larger vertex IDs (or vice versa) to be able to count each triangle once. In the block-based context, this rule can be applied by considering subgraph indices in increasing order, such that $i \leq j \leq k$. With this restriction, if the vertex set, V , of a given graph partitioned into p parts, then number of tasks, $|T|$, is defined as:

$$|T| = \frac{p \times (p+1) \times (p+2)}{6}$$

Algorithm 4: BBTaskComposition (G, p)

```

 $T \leftarrow \emptyset$                                  $\triangleright$  Initialize empty task list
for  $i = 0$  to  $p - 1$  do
  for  $j = i$  to  $p - 1$  do
    for  $k = j$  to  $p - 1$  do
       $T \leftarrow T \cup \{G_{i,j}, G_{j,k}, G_{i,k}\}$ 
return  $T$ 

```

For instance, as shown in the toy example presented in Fig. 2 10, tasks can be defined. Alg. 4 illustrates task list composition for a given symmetric rectilinear partitioned graph.

4.3 Counting triangles in a task**Algorithm 5:** BBTC-INTERSECT ($G_{i,j}, G_{j,k}, G_{i,k}$)

```

 $\tau \leftarrow 0$                                  $\triangleright$  Initialize number of triangles to 0
for each  $u \in V_S(G_{i,j})$  do
  for each  $v \in N(G_{i,j}, u)$  do
     $\tau \leftarrow \tau + \text{NINTERSECT}(N(G_{i,k}, u), N(G_{j,k}, v))$ 
return  $\tau$ 

```

In the context of this paper a task is defined as counting the number of triangles that appear in three subgraphs; $G_{i,j}$, $G_{j,k}$ and $G_{i,k}$ where $i \leq j \leq k$. For each edge, (u, v) in $G_{i,j}$, this operation can be done by counting the number of common neighbors between the partial neighbor list of u in $G_{i,k}$ ($N(G_{i,k}, u)$) and the partial neighbor list of v in $G_{j,k}$ ($N(G_{j,k}, v)$). Similar to regular triangle counting algorithms this can be computed using list or hashmap-based intersection algorithms as described in Alg. 5 and Alg. 6 respectively.

Algorithm 6: BBTC-HMAP ($G_{i,j}, G_{j,k}, G_{i,k}$)

```

 $\tau \leftarrow 0$                                  $\triangleright$  Initialize number of triangles to 0
for each  $u \in V_S(G_{i,j})$  do
  for each  $v \in N(G_{i,k}, u)$  do
     $H[v] \leftarrow u$ 
    for each  $w \in N(G_{j,k}, v)$  do
      if  $H[w] = u$  then
         $\tau \leftarrow \tau + 1$ 
return  $\tau$ 

```

4.4 Computational Complexity

In this subsection, we will present the complexity analysis of BBTC. Let the input graph $G = (V, E)$ be partitioned with a $p \times p$ symmetric rectilinear partitioning. Let m_{\max} and m_{avg} be, respectively, the maximum and average number of edges (nonzeros) within subgraphs (blocks). We also define *load imbalance*, λ , as $\lambda = \frac{m_{\max}}{m_{\text{avg}}}$. Let d_{\max} be the maximum degree in the input graph, i.e.,

$$d_{\max} = \max_{u \in V} d(G, u),$$

and d'_{\max} be the maximum degree within all blocks, i.e.,

$$d'_{\max} = \max_{u \in V_S(G_{i,j}), \forall 0 \leq i, j < p} d(G_{i,j}, u).$$

(Please note that, they can be degrees of different vertices). We define c as the ratio of these max degrees, i.e., $c = \frac{d_{\max}}{d'_{\max}}$. Clearly $1 \leq c \leq p$.

One may see Alg. 5 and Alg. 6 as iterating over the edges in $G_{i,j}$ and counting common neighbors of source and destination vertices' partial neighbor lists that appear in $G_{i,k}$ and $G_{j,k}$, respectively. Hence, since there can be at most m_{\max} edges in $G_{i,j}$ and maximum degree of a vertex in a block is defined with d'_{\max} , a single task can be solved in $O(m_{\max}d'_{\max})$ time. The maximum number of nonzeros (m_{\max}) among the blocks can be defined using load imbalance as, $m_{\max} = \lambda m_{\text{avg}}$. There are $\frac{p(p+1)}{2}$ blocks, hence the average number of nonzeros (m_{avg}) per block is $m_{\text{avg}} = \frac{2m}{p(p+1)}$. When we replace m_{avg} with that representation, m_{\max} equals to $\lambda \frac{2m}{p(p+1)}$. A task can be solved in

$$O\left(\lambda \frac{2m}{p(p+1)} \frac{d_{\max}}{c}\right)$$

time. Since there can be at most

$$\frac{p(p+1)(p+2)}{6}$$

concurrent tasks, our block-based triangle counting problem can be solved in

$$O\left(\frac{p(p+1)(p+2)}{6} \lambda \frac{2m}{p(p+1)} \frac{d_{\max}}{c}\right)$$

which is equal to

$$O\left(\frac{\lambda p}{c} m d_{\max}\right)$$

when we apply proper simplifications. When the vertices are ordered based on their degrees, the intersection operation between neighbor lists of two vertices can be computed in $O(\sqrt{m})$ [1] instead of $O(d_{\max})$. Hence, block-based triangle counting, using the degree ordering of vertices, can be solved in

$$O\left(\frac{\lambda p}{c} m^{3/2}\right).$$

Note that, for dense graphs, achieving the perfect load imbalance ($\lambda \approx 1$) is straightforward, and with high probability, c and p would be nearly equal. Hence, in dense instances, the computational complexity of the block-based formulation is the same as the work-optimal triangle counting algorithms. In irregular problem instances, the load imbalance is crucial in the computational complexity. Therefore, balanced partitioning of the graph is essential. When the partitioned adjacency matrix has the perfect load imbalance, the computational complexity of the irregular problem instances becomes nearly equal to the work-optimal cases.

5 PARALLEL AND HETEROGENEOUS EXECUTION

The BBTC algorithm can be parallelized by concurrently executing the tasks. The final number of triangles is the sum of the triangles found on all the tasks. In this work, we utilize CPUs and GPUs to process all the tasks. At a high level, there are no architecture-specific changes in the algorithm. That allows the execution of our algorithm on any accelerator and CPU combination. However, task handling differs between CPU and GPU because of architectural differences. While a CPU thread executes a single task,

threads execute cooperatively to complete a task in parallel on GPUs. Alg. 9 illustrates this difference.

In hybrid execution environments, assigning computationally heavy tasks to GPUs and lighter tasks to CPUs (e.g., [29]) is one of the essential optimizations to leverage the massive parallelism available on GPUs. We implement a similar approach. Tasks are ordered based on their estimations in the task queue (Q). GPUs start with heavy tasks from one end of the execution queue, and CPUs process light tasks from the other end of the execution queue. They continue to move towards each other as they complete the assigned tasks. BBTC keeps a global pointer, (q_c), for the last task id that a CPU thread starts to execute. Each CPU thread atomically decrements the pointer until finding an unprocessed task or terminates if the counter reaches the cut-off point of the task queue. That continues while the execution queue is not empty. Fig. 3 illustrates this procedure. The execution of tasks by GPUs and CPUs are described in Alg. 7 and Alg. 8, respectively.

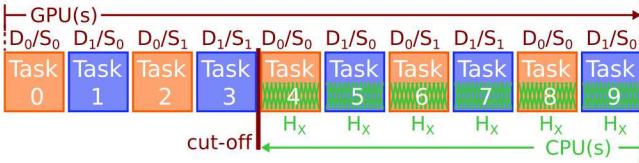


Fig. 3. Task Scheduling Between Devices and Host. D_i : i^{th} Device, S_i : i^{th} Stream on a Device, H_x : a CPU

For sorting, we estimate the workload of a task, t , using:

$$ExecTime(t) = |\{(u, v) \in G_{i,j}\}| \times \max\{\delta(G_{i,k}), \delta(G_{j,k})\}$$

where $\delta(G_{i,j})$ represents the average degree of a subgraph $G_{i,j}$ i.e.,

$$\delta(G_{i,j}) = \frac{\sum_{u \in V_i} d(G_{i,j}, u)}{|V_i|}.$$

We sort tasks in the non-increasing order based on their estimation weights. To decrease the initial synchronization cost of GPUs and guarantee the assignment of heavier tasks to the GPUs, a cut-off (red line in Figure 3) is pre-defined. CPUs do not go past the cut-off.

We use CUDA streams to execute several tasks on the GPUs simultaneously. We create four CUDA streams, ($n_s = 4$), for one of the n_g GPUs. Then, one of the n_c CPU threads is assigned to a stream (see Alg. 9). That CPU thread is responsible for waiting on that stream, synchronizing the stream, sending a task to a device through that stream, and gathering information from a device through that stream. All of these operations use asynchronous function calls. When we create streams and assign a thread for each of them, GPUs and CPUs compete for tasks and get a new one from the queue when they finish executing a task.

The cut-off allows all GPU streams to run tasks until the cut-off without worrying about the next assignment (the first while loop in Alg. 7). GPUs can go past the cut-off line if more tasks are available (the second while loop in Alg. 7).

We allow GPUs and CPUs to compete for the tasks. However, GPUs do not compete among themselves. We assign tasks to GPUs in a deterministic round-robin fashion. This approach allows stream threads to know the next task

Algorithm 7: RUNTASKONGPU (tId)

```

 $gId \leftarrow tId \bmod n_s$                                 ▷ GPU id
 $sId \leftarrow tId/n_s$                                      ▷ Stream id
while  $tId < cut-off$  do
  for each  $G_{i,j} \in T[tId]$  do
    if not ISCOPIED ( $G_{i,j}, gId$ ) then
      ASYNCOPY ( $G_{i,j}, gId$ )
     $G_{i,j}, G_{j,k}, G_{i,k} \leftarrow T[tId]$       ▷ Get copied blocks
    if ISDENSE ( $T[tId]$ ) then
      async BBTC-HMAP ( $G_{i,j}, G_{j,k}, G_{i,k}$ )
    else
      async BBTC-INTERSECT ( $G_{i,j}, G_{j,k}, G_{i,k}$ )
     $tId \leftarrow tId + n_g \times n_s$ 
  Copy blocks for next task,  $tId$ .
for each  $G_{i,j} \in T[tId]$  do
  if not ISCOPIED ( $G_{i,j}, gId$ ) then
    ASYNCOPY ( $G_{i,j}, gId$ )
  Get a task from  $Q$  if available, stop otherwise
while  $tId < |T|$  and not ATOMICSET ( $Q, tId$ ) do
  SYNCSTREAM ( $sId, gId$ )          ▷ Synchronize the
  stream
   $G_{i,j}, G_{j,k}, G_{i,k} \leftarrow T[tId]$       ▷ Get copied blocks
  if ISDENSE ( $T[tId]$ ) then
    async BBTC-HMAP ( $G_{i,j}, G_{j,k}, G_{i,k}$ )
  else
    async BBTC-INTERSECT ( $G_{i,j}, G_{j,k}, G_{i,k}$ )
   $tId \leftarrow tId + n_g \times n_s$ 
  for each  $G_{i,j} \in T[tId]$  do
    if not ISCOPIED ( $G_{i,j}, gId$ ) then
      ASYNCOPY ( $G_{i,j}, gId$ )

```

Algorithm 8: RUNTASKONCPU ()

```

 $tId \leftarrow$  ATOMICDECREMENT ( $q_c$ )      ▷ Get task index.
 $\tau \leftarrow 0$ 
while  $t \geq cut-off$  do
  Check if task has been processed and set if not.
  if not ATOMICSET ( $Q, tId$ ) then
     $G_{i,j}, G_{j,k}, G_{i,k} \leftarrow T[tId]$ 
    if ISDENSE ( $T[tId]$ ) then
       $\tau \leftarrow \tau +$  BBTC-HMAP ( $G_{i,j}, G_{j,k}, G_{i,k}$ )
    else
       $\tau \leftarrow \tau +$  BBTC-INTERSECT
      ( $G_{i,j}, G_{j,k}, G_{i,k}$ )
    else
       $tId \leftarrow$  ATOMICDECREMENT ( $q_c$ )
  return  $\tau$ 

```

that can be executed by that stream. Then a stream thread can overlap copying the blocks of the next assignment with the computation.

6 RELATED WORK

Inspired by *AYZ listing* [8], Latapy [1] proposed a refinement technique that does not require $\theta(n^2)$ space. Latapy's algorithm uses the list-based intersection for small degree vertices and the hashmap-based intersection for high degree vertices. In this work, we use a similar approach; i.e., any

Algorithm 9: BBTC (G, p)

```

▷ Initialization of the system variables.
 $n_c \leftarrow \text{NCORES}(); n_g \leftarrow \text{NGPUS}(); n_s \leftarrow 4;$ 
 $\tau \leftarrow 0$                                 ▷ Number of triangles

▷ Partitioning
 $G(V^p, E) \leftarrow \text{PBD}(G, p)$     ▷ Parallel PBD algorithm

▷ Composing task list
 $T \leftarrow \text{BBTASKCOMPOSITION}(G, p)$     ▷ Composing
    task list
 $T \leftarrow \text{ORDERTASKS}(T)$  ▷ Heuristic based ordering of
    tasks

▷ Initializing a list of bit flags for each task
 $Q[i] = 0$ , for  $0 \leq i \leq |T| - 1$ 
 $q_c \leftarrow |T|$                                 ▷ Queue index for competing cores.

▷ Spawning gpu threads: one per each stream
for  $i = 0$  to  $n_s - 1$  do
  for  $j = 0$  to  $n_g - 1$  do
     $\quad \quad \quad$  spawn RUNTASKONGPU ( $i \times s + j$ )

▷ Spawning cpu threads
for  $i = n_s \times n_g$  to  $n_c$  do
   $\quad \quad \quad$  spawn RUNTASKONCPU ()

▷ Gathering & aggregating counts
for  $i = 0$  to  $n_s - 1$  do
  for  $j = 0$  to  $n_g - 1$  do
     $t \leftarrow \text{GETCOUNT}(i, j)$ 
    ATOMICADD ( $\tau, t$ )
return  $\tau$ 

```

task with sparse blocks uses the list-based intersection, and any task consist of dense blocks uses a hashmap. Shun and Tangwongsan [2] parallelized Latapy’s *compact-forward* algorithm. Recently, Dhulipala et al. [4] re-implemented an optimized version of this algorithm.

A linear-algebra-based triangle counting implementation, kkTri (previously designated TCKK), was proposed by Wolf et al. [30]. That work focused on efficient shared memory parallelism on top of a portable SpGEMM (called KK-MEM) [31] in the Kokkos Kernels library¹. This work optimized two linear-algebra-based formulations of the triangle counting problem and used different types of hashmap accumulators based on the sparsity of the graphs. In a later work, Yaşar et al. [18] improved this algorithm by using a 1D task-based approach and an adaptive runtime. In this paper, we use 2D partitioning instead of 1D partitioning, and we always use dense hashmaps. The memory requirement for dense hashmaps decreases significantly because of the 2D partitioning. The smaller dense hashmaps are also more suitable for GPU architectures.

Green et al. [17] proposed a list-intersection-based algorithm on a single GPU. In that work, neighbor lists of source and destination vertices of each edge are concurrently processed by 32 threads from a warp. However, this approach comes with expensive partitioning overhead. Hu et al. [3] try to overcome this problem by introducing a binary search-based intersection method. Both works copy the graph into

TABLE 2
Overview of the Architectures. Pinned: transfers through pinned memory. Pageable: transfers through pageable memory. Peer to peer: transfers done between devices. H2D: host to device. D2H: device to host. (measured bandwidth)

	DGX	Newell	Haswell
CPU	Intel, E5-2698	POWER9	Intel, E7-4850
Cores	2×20	2×16	4×14
Host Memory	512 GB	320 GB	2 TB
L2-Cache	256 KB	512 KB	256 KB
L3-Cache	50 MB	10 MB	35 MB
GPU	V100	V100	N/A
GPU Memory	32 GB	32 GB	N/A
Number of GPUs	8	2	N/A
Pageable	H2D D2H	9.2 GB/s 8.0 GB/s	12.2 GB/s 14.2 GB/s
Pinned	H2D D2H	10.7 GB/s 12.1 GB/s	60.0 GB/s 60.0 GB/s
Peer to Peer		23.5 GB/s	31.3 GB/s

GPU memory before starting the computation, which causes significant efficiency issues because of the idle time. Besides, using that approach [17], one cannot process graphs that are larger than the GPU’s memory size. [3] addresses this problem by applying a rectilinear partitioning strategy; however, their execution time is drastically affected by the number of partitions. Date et al. [15] proposed to use CPUs and GPUs in a collaborative fashion using GPU zero-copy memory, aiming to decrease CPU-GPU data transfer overhead by leveraging unified memory capabilities. Inefficient use of GPU memory causes drastic performance decrease in that case.

Recently, Tom et al. [32] and Acer et al. [33] proposed distributed-memory triangle counting algorithms. Tom et al. [32] propose a 2D distributed-memory triangle counting algorithm that follows similar steps to Cannon’s parallel matrix multiplication algorithm. Acer et al. [33] propose an MPI based distributed memory parallel 2D triangle counting algorithm which tries to exploit the benefits of shared memory parallelism using Cilk.

7 EXPERIMENTAL EVALUATION

We present several experiments to identify the performance trade-offs of the proposed work. In our tests, we used three architectures with multi-core processors and multi-GPUs. Table 2 lists the properties of those architectures. Depending on the architecture GNU compiler (g++) version 7.2 or Intel compiler version 19.03 (icpc), CUDA runtime version 10.0 and OpenMP version 4.0 are used to compile and run the code. The source code of BSD-licensed BBTC is available at <http://tda.gatech.edu/software/bbtc/>.

7.1 Used state-of-the-art works for comparison

In our experiments, we compared BBTC’s performance with three state-of-the-art implementations: (1) TCM², a multi-core triangle counting algorithm proposed by Shun and

1. <https://github.com/kokkos/kokkos-kernels>

2. TCM: <https://github.com/Ldhulipala/gbbs>

TABLE 3

Overview of Design Choices. [†] Complexity for one dimensional parallel case. Complexity of partitioned version is not provided in [3].

	State of The Art			
	TCM	kkTri	TriCore	BBTC
Algorithmic properties				
List Intersect.	✓	✗	✗	✓
Map Intersect.	✗	✓	✗	✓
Search Intersect.	✗	✗	✓	✗
Multi-Core	✓	✓	✗	✓
Multi-GPU	✗	✗	✓	✓
Complexity	$O(m^{\frac{3}{2}})$	$O(m^{\frac{3}{2}})$	$O(m^{\frac{3}{2}} \log \sqrt{m})$ [†]	$O(\frac{\lambda p}{c} m^{\frac{3}{2}})$
Implemented optimizations				
Vertex Ordering	✓	✓	✓	✓
Compression	✓	✓	✗	✗
CUDAStreams	N/A	N/A	✗	✓
Parallelization strategy				
One-dimension	✓	✓	✓	✗
Two-dimension	✗	✗	✓	✓
Used runtimes for implementation				
Pthread based	✓	✗	✗	✗
OpenMP	✓	✓	✗	✓
Cilk	✓	✓	✗	✓
TBB	✗	✗	✗	✓
CUDA	✗	✗	✓	✓

Tangwongsan [2] and optimized by Dhulipala et al. [4]. We use Dhulipala et al. [4]’s optimized version. (2) kkTri ³, a multi-core linear algebra-based triangle counting algorithm proposed by Wolf et al. [30] and improved by Yaşar et al. [18]. We use the optimized version on Intel architectures. (3) TriCore ⁴, a multi-GPU triangle counting algorithm implemented by Hu et al. [3] and improved in [34]. We use TriCore’s optimized (latest) version [34]. While TCM and kkTri algorithms have the optimal complexity, TriCore and BBTC algorithms have higher complexities due to their design choices. Note that, the complexity of TriCore is for the one-dimensional parallel case. TriCore also implements a variant of rectilinear partitioning [24] to be able to process graphs that cannot fit into GPU memory. This partitioning increases the cost of their algorithm. However, the complexity of TriCore’s partitioned version is not provided in [3]. Table 3 presents a high-level overview of the design choices of these algorithms.

7.2 Dataset and peak rates

We evaluated the BBTC algorithm on 16 real-world, and 8 synthetic graphs (RMAT) from the SuiteSparse Matrix Collection [35], SNAP ⁵, web data commons ⁶, and GraphChallenge 2014-04

large ⁷. Table 4 presents, the properties of these graphs: The number of vertices ($|V|$), the number of edges ($|E|$), the number of triangles, the compression ratio of the graph (*Comp.*), the size of the graph in memory (Raw Size), the number of tiles, and the number of tasks. For all experiments, we get the median of five runs with different numbers of GPUs and report the best. For a graph, the rate is defined as the ratio between the execution time and the number of edges. Best rates are achieved on the DGX machine. Table 4 also reports rates for each graph.

7.3 Sequential execution time for varying partitions

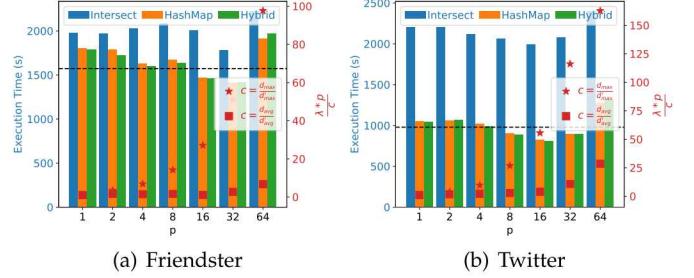


Fig. 4. Sequential time of Friendster and Twitter graphs for different p values. Black dashed line represents Latapy’s algorithm’s execution time.

Figure 4 shows the sequential behavior of the BBTC algorithm with respect to the different number of partitions ($p \times p$) on the Friendster and the Twitter graphs. Each bar represents the execution time (left-axis) of the BBTC algorithm using the list-based intersection, hashmap-based intersection and hybrid intersection, for a given p (x-axis). The black dashed line represents the best sequential time that we get using Latapy’s algorithm [1]. We observe that in both graphs when p is increased, sequential execution time first decreases due to better memory utilization and then increases because of higher load imbalance. Note that, the BBTC algorithm runs in $O(\frac{\lambda p}{c} m^{\frac{3}{2}})$ time. To analyze the behaviour of the BBTC algorithm, we have plotted $\frac{\lambda p}{c}$ (right-axis) values for $c = \frac{d_{\max}}{d_{\max}}$ (star-shaped marker) and $c = \frac{d_{\max}}{d_{\max}}$ (square-shaped marker) where d_{\max} , d_{\max} are the maximum and the average degrees in the given graph, and d'_{\max} , d'_{\max} are the maximum and the average degrees in the blocks of the partitioned graph, respectively. We observe that trend in execution times are more co-related with $c = \frac{d_{\max}}{d'_{\max}}$. Hence, one should pick p around the average degree of the graph. Note that, the BBTC algorithm outperforms Latapy’s algorithm on both graphs for p values closer to d_{\max} .

7.4 Comparison with Latapy’s sequential algorithm

Given K , Latapy’s algorithm [1] applies hashmap-based intersection for vertices whose degrees are higher than K , and list-based intersection for the others. We evaluate Latapy’s algorithm’s [1] performance with respect to different K values. Fig. 5(a) illustrates the performance profile of the algorithm. In the performance profile, we plot the number

3. kkTri: <https://github.com/Kokkos/kokkos-kernels>

4. TriCore: <https://github.com/huyang1988/TC>

5. SNAP Datasets: <http://snap.stanford.edu/data>

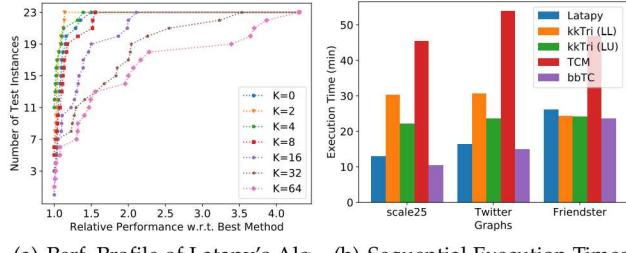
6. WDC Dataset: <http://webdatacommons.org/hyperlinkgraph/>

7. Graph Challenge Datasets: <https://graphchallenge.mit.edu/data-sets>

TABLE 4

Properties of the dataset. Best of the medians of execution times in seconds and corresponding rates are reported.

Data Set	$ V $	$ E $	Triangle Count	Comp.	Raw Size	Tiles	Tasks	Best (Copy Included) Time (s)	Rate ($\times 10^8$)
cit-HepTh	27,770	352,285	1,478,735	0.6	2.2 MB	64	120	0.002	1.7
email-EuAll	265,214	364,481	267,313	0.5	9.5 MB	64	120	0.002	1.9
soc-Epinions1	75,879	405,740	1,624,481	0.6	3.9 MB	64	120	0.002	2.1
cit-HepPh	34,546	420,877	1,276,868	0.5	2.7 MB	64	120	0.002	2.0
soc-Slashdot0811	77,360	469,180	551,724	0.6	5.4 MB	144	364	0.002	2.9
soc-Slashdot0902	82,168	504,230	602,592	0.6	5.7 MB	144	364	0.002	3.1
flickrEdges	105,938	2,316,948	107,987,357	0.6	14 MB	144	364	0.016	1.5
amazon0312	400,727	2,349,869	3,686,467	0.7	28 MB	144	364	0.006	3.9
amazon0505	410,236	2,439,437	3,951,063	0.7	29 MB	144	364	0.007	3.7
amazon0601	403,394	2,443,408	3,986,507	0.7	28 MB	144	364	0.006	4.4
scale18	174,147	3,800,348	82,287,285	0.6	26 MB	256	816	0.021	1.8
scale19	335,318	7,729,675	186,288,972	0.6	56 MB	400	1540	0.041	1.9
as-Skitter	1,696,415	11,095,298	28,769,868	0.8	146 MB	400	1540	0.027	4.1
scale20	645,820	15,680,861	419,349,784	0.6	110 MB	400	1540	0.079	2.0
cit-Patents	3,774,768	16,518,947	7,515,023	0.5	352 MB	400	1540	0.038	4.4
scale21	1,243,072	31,731,650	935,100,883	0.6	216 MB	400	1540	0.144	2.2
soc-LiveJournal1	4,847,571	42,851,237	285,730,264	0.7	534 MB	400	1540	0.121	3.6
scale22	2,393,285	64,097,004	2,067,392,370	0.6	464 MB	576	2600	0.325	2.0
scale23	4,606,314	129,250,705	4,549,133,002	0.5	915 MB	576	2600	0.549	2.4
scale24	8,860,450	260,261,843	9,936,161,560	0.5	1.9 GB	784	4060	1.154	2.3
scale25	17,043,780	523,467,448	21,575,375,802	0.5	4.0 GB	1024	5984	2.400	2.2
twitter	61,578,414	1,202,513,046	34,824,916,864	0.5	13 GB	1296	8436	4.582	2.6
friendster	65,608,366	1,806,067,135	4,173,724,142	0.5	16 GB	1296	8436	3.133	5.8
wdc-2014	1,724,573,718	124,141,874,032	4,587,563,913,535	0.8	650 GB	1296	8436	116.5	12.4

Fig. 5. Evaluation of Latapy's algorithm wrt. K (5(a)) and comparison of sequential execution times (5(b)).

of the test instances (y-axis) in which a K value obtains an execution time on an instance that is no larger than x times (x-axis) the best execution time achieved by any K value for that instance [36]. Therefore, the higher a profile at a given x value, the better a K value is. We observe that switching between list-based and hashmap-based intersection techniques improves efficiency. In the majority of the test instances, $K = 2$ performs better than others due to the number of smaller graphs in our dataset. In larger instances, such as Twitter and Friendster, $K = 32$ and $K = 64$ gives the best performance. Hence, K should be picked based on the graph size.

Fig. 5(b) illustrates sequential execution times of Latapy's algorithm, kkTri with LL, and LU formulations [30],

TCM [4], and BBTC on the three large graphs of our dataset; scale25, Twitter and Friendster. BBTC outperforms other algorithms in all graph instances due to better memory utilization. TCM performs the worst because of the list-based intersection. TCM achieves better scalability in parallel settings with list-based intersection, because using dense hashmap (i.e. an array of the length $n = |V|$) causes poor memory utilization. However, in a sequential setting list-based intersection performs worse than hashmap-based intersection because there is no bandwidth competition among running threads. kkTri uses a linked-list-based hashmap for large graphs, which helps it to outperform Latapy's algorithm on the Friendster graph. We observe that kkTri's LU formulation outperforms LL formulation in the sequential case while LL performs better in parallel setting [18].

7.5 Task Workload Estimation

Fig. 6(a) and Fig. 6(b) present execution times of the most time consuming 32 tasks (sorted on the x-axis) on a CPU and a GPU for Friendster and Twitter graphs. We observe that a GPU executes heavy tasks at least seven times faster than a CPU. Hence, assigning heavy tasks to GPUs is crucial. Note that, BBTC doesn't need to estimate workload of tasks with full accuracy, just identifying heavy tasks is enough to increase GPU utilization. Therefore, using a thorough model to estimate the workload is an overkill. In this experiment,

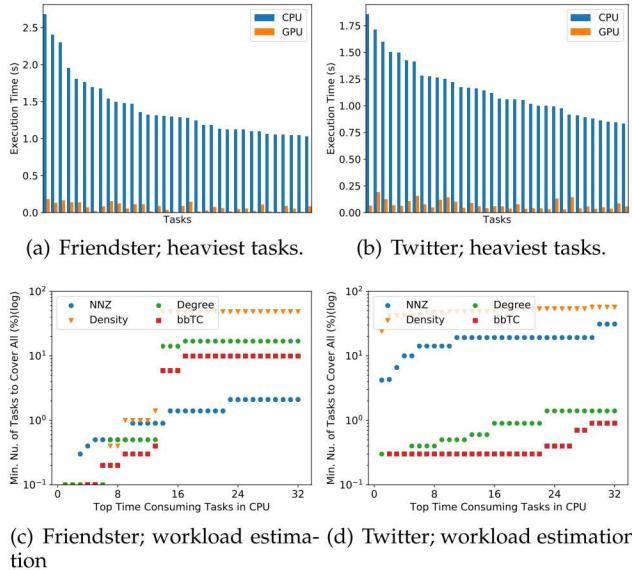


Fig. 6. Comparison of Different Estimation Functions

we evaluate the effectiveness of our proposed workload estimation heuristic (see Sec. 5).

Fig. 6(c) and Fig 6(d) present the least required percentage of tasks (y-axis in log-scale) that starts from the beginning of a sorted task list using an estimation function, to cover most x time-consuming tasks (x-axis) (lower is better). In this experiment, we compare four different estimation functions. NNZ: the total number of nonzeros in a task. Density: the sum of the average nonzeros per unit in the blocks of a task. Degree: the sum of the average degree of blocks in a task and BBTC's estimation function that we have described in Sec. 5. We observe that BBTC's estimation heuristic outperforms the others in 47 of 64 instances. In the Friendster graph, NNZ outperforms BBTC's estimation heuristic in predicting the most time consuming 17 to 32 tasks. However, on the Twitter graph, NNZ performs poorly. Hence, using NNZ as an estimation function may end up with unstable performance. In both graphs, not surprisingly, degree-based estimation performs closer to BBTC's estimation heuristic due to the effect of vertex degrees on common neighbor computation.

7.6 Comparison with TriCore

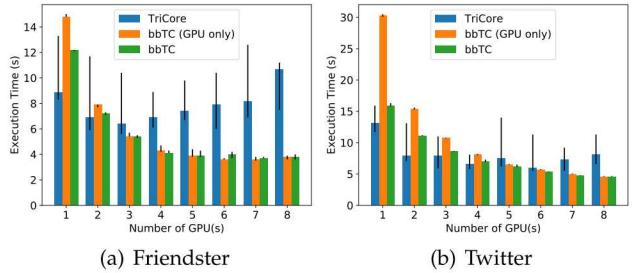


Fig. 7. Comparison with TriCore

In this experiment, we compare BBTC with TriCore using Friendster and Twitter graphs. Note that, partitioning is

turned off for TriCore as it affects the performance negatively. For Twitter and Friendster graphs we can use TriCore without partitioning because both graphs can fit into GPUs' memory. However, when one needs to run larger problem instances, partitioning is going to be necessary and will hurt TriCore's performance due to bad memory utilization. Each bar in Fig. 7 illustrates the execution time (y-axis) of an algorithm for different number of GPUs (x-axis). We observe that BBTC scales better than TriCore with respect to the number of GPUs thanks to its efforts to overlap communication with computation. TriCore's scalability becomes worse after 4 GPUs on Friendster, and its times go up after 4 GPUs on Twitter. On the other hand, BBTC scales much better up to 5 GPUs and continues to improve slightly, after 5 GPUs. In Fig. 7, each error-bar presents variation between minimum and maximum execution times among five runs. We observe that the TriCore algorithm is highly unstable, and its runtime deviates up to 40% while this deviation is less than 5% for BBTC. TriCore's high deviation might be caused by unstable warp usage during intersection computation.

Symmetric rectilinear partitioning may generate sparse blocks due to the irregularity of the graphs and also a generalization of the partitioning. Tasks that contain highly sparse blocks might be executed extremely fast by CPUs and GPUs. However, processing these tasks on GPUs, come with an overhead caused by data copy, API calls, and stream synchronization costs. Hence, assigning all tasks on GPUs may end up with poor performance and also contradicts with BBTC's design goals. BBTC targets heterogeneous execution. As shown in Fig. 7 because of the overhead of the light tasks in 1 and 2 GPU-only cases, TriCore outperforms BBTC. Hybrid execution addresses this issue and BBTC achieves a competitive performance in these cases. Note that, even in GPU only case, BBTC scales better than TriCore and outperforms it after 4 GPUs in the system.

7.7 Overlapping communication and computation

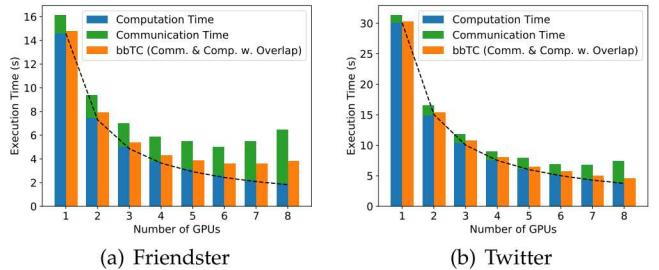


Fig. 8. Overlap comparison. Dashed line represents linear speedup.

Fig. 8 illustrates how successfully BBTC overlaps communication with computation in the different number of GPU settings. In Fig. 8, blue bars represent computation only execution time (y-axis), and green bars represent communication time between CPUs and GPUs. Orange bars represent BBTC's GPU only execution times, which tries to overlap communication with computation. The black dashed line represents the linearly scaled execution time of single GPU computation-only time for the number of

GPUs. As expected, BBTC’s computation time scales almost linearly with the number of GPUs. We observe that on the Friendster graph, the BBTC algorithm can overlap communication with computation with just 10% overhead up to 5 GPUs, however this overhead increases for the number of GPUs and becomes 50% when the number of GPUs is 8, due to bandwidth limitations. This overlap varies from 1% to 18% on the Twitter graph.

7.8 Hybrid execution and Cut-off

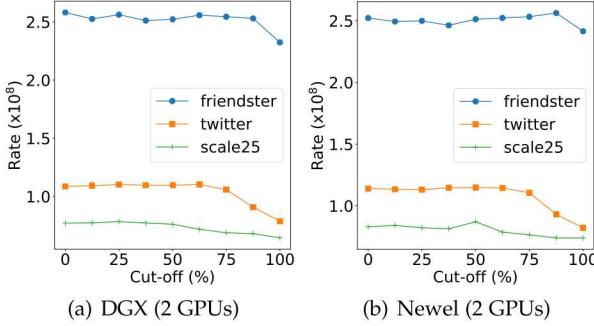


Fig. 9. Hybrid execution and cut-off.

In this experiment, we evaluate the effect of the cut-off location (see Fig. 3) from 0 to $|T|$. If the cut-off is equal to the number of tasks ($|T|$), we assign all the tasks to the GPUs. If the cut-off is 0, then CPUs may execute all the tasks. In this experiment, we set cut-off as a multiple of one eighth of the number of tasks ($\frac{|T|}{8}$), and report rates for scale25, Twitter, and Friendster graphs on DGX and Newell architectures. As illustrated in Fig. 9 when the cut-off is larger than the last two quartiles, then the overall performance decreases because the execution of the lighter tasks on GPUs doesn’t compensate the overhead. The cut-off creates a one-way barrier and guarantees the assignment of heavy tasks on GPU by restricting CPUs for not passing the cut-off point. Hence, the cut-off has crucial importance. However, thanks to estimation function and relatively fewer heavy tasks deciding the cut-off point is not a hard problem. By default, the cut-off point is the middle of the execution queue.

7.9 Relative speedup

Figure 10 presents relative speedup between this work and two state-of-the-art algorithms; TCM [2], [4], and kkTri [30]. We compare BBTC with these two state-of-the-art works for CPU only, and heterogeneous scenarios. In the CPU only scenario, BBTC outperforms TCM in 23 of 23 cases and kkTri in 20 of 23 cases. kkTri can perform better than BBTC in three small instances; email-EuAll, amazon0505 and amazon0611. BBTC can achieve up to 18 \times and 14 \times speedup with respect to TCM and kkTri, respectively. In the heterogeneous case, BBTC outperforms TCM and kkTri in 23 of 23 cases. However, in four small instances (email-EuAll, amazon0505, amazon0611, and cit-Patents), TCM and BBTC are comparable. BBTC can achieve up to 18 \times speedup on large graphs and up to 22 \times speedup on smaller instances.

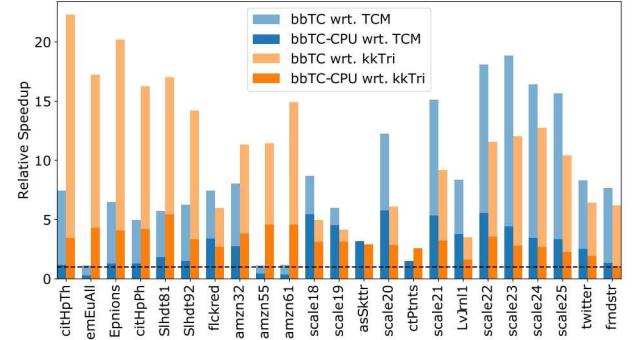


Fig. 10. Relative speedup between this work and TCM and kkTri algorithms, on Newell architecture. bbTC-CPU: speedup when we only use CPUs. bbTC: speedup when we use CPUs and GPUs. Black dashed line represents the baseline. Graphs are sorted on x-axis based on their number of edges.

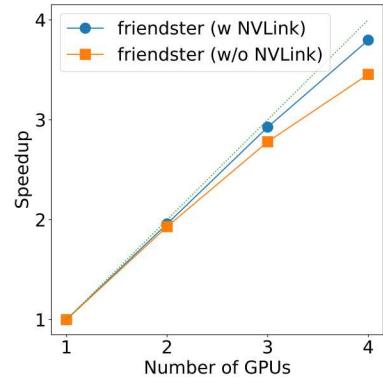


Fig. 11. Effect of bandwidth.

7.10 Effect of bandwidth on speedup

BBTC assigns four CUDA streams per GPU. Each of these streams asynchronously copies the data from the host memory to the device memory. The bandwidth between the host machine and the device becomes a bottleneck when there are more GPUs on the node. In this experiment, we evaluate how the bandwidth can affect the scalability. Fig. 11 illustrates BBTC’s strong scalability on Friendster graph up to four GPUs on a machine with NVLink and a machine without NVLink. As expected, with NVLink enabled architecture we get better scaling. Hence, if we would have a server with NVLink and 8 Volta GPUs with, BBTC’s could get even better execution times.

7.11 Comparison on the WDC-2014 graph

TABLE 5
Execution times of four different algorithms on two different architectures on a WDC graph.

	Haswell	DGX
kkTri	423	Out of memory
TCM	476	Out of memory
TriCore	N/A	
bbTC	265	116

In this experiment, we evaluate the performance of the BBTC algorithm on a WDC hyperlink graph. We present runtimes in seconds in Table 5. This graph has 1.7 billion vertices and 124 billion edges. DGX server doesn't have enough memory to execute this graph. Therefore, kkTri and TCM algorithms failed on DGX. TriCore also raised an exception during its partitioning stage; consequently, we couldn't run TriCore. To be able to process such a graph, which is larger than available memory, BBTC uses memory-mapped files. In CPU only execution (Haswell) BBTC outperforms kkTri and TCM 1.6 \times and 1.8 times, respectively. BBTC can process the WDC-2014 graph in 116 seconds on DGX. Note that, on DGX a hardware RAID system is installed. Hence, BBTC can perform faster disk I/O.

8 CONCLUSIONS

We have developed a triangle counting algorithm called BBTC, which leverages a special 2D partitioning scheme called symmetric rectilinear partitioning to define tasks that can be easily used on any heterogeneous accelerator, CPU combination. To leverage from the massive computing capabilities of the GPUs and to overlap communication with computation, BBTC implements a dynamic task-scheduling scheme that schedules tasks on multi-core-CPUs and multiple streams on multiple GPUs. BBTC is up to 20 \times faster than state-of-the-art multi-core CPU algorithms and 1.5 \times faster than state-of-the-art multi-GPU algorithm. Our experimental results demonstrate that BBTC achieves the fastest end-to-end execution times when the graph is not on the GPU.

Acknowledgments: This work was supported in parts by the NSF grants CCF-1919021 and Sandia National Laboratories/Sandia Corp contract 2115504. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

REFERENCES

- [1] M. Latapy, "Main-memory triangle computations for very large (sparse (power-law)) graphs," *Theoretical Computer Science*, pp. 458–473, 2008.
- [2] J. Shun and K. Tangwongsan, "Multicore triangle computations without tuning," in *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 149–160.
- [3] Y. Hu, H. Liu, and H. H. Huang, "Tricore: Parallel triangle counting on gpus," in *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2018, pp. 171–182.
- [4] L. Dhulipala, G. E. Blelloch, and J. Shun, "Theoretically efficient parallel graph algorithms can be fast and scalable," in *Proceedings of the 30th on Symposium on Parallelism in Algorithms and Architectures*. ACM, 2018, pp. 393–404.
- [5] J. W. Berry, L. K. Fostvedt, D. J. Nordman, C. A. Phillips, C. Seashadri, and A. G. Wilson, "Why do simple algorithms for triangle enumeration work in the real world?" in *Theoretical Computer Science*, 2014, pp. 225–234.
- [6] A. Itai and M. Rodeh, "Finding a minimum circuit in a graph," *SIAM Journal on Computing*, vol. 7, no. 4, pp. 413–423, 1978.
- [7] M. Ortmann and U. Brandes, "Triangle listing algorithms: Back from the diversion," in *Proceedings of the Meeting on Algorithm Engineering & Experiments*. SIAM, 2014, pp. 1–8.
- [8] N. Alon, R. Yuster, and U. Zwick, "Finding and counting given length cycles," *Algorithmica*, vol. 17, no. 3, pp. 209–223, 1997.
- [9] R. Pagh and F. Silvestri, "The input/output complexity of triangle enumeration," in *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2014, pp. 224–233.
- [10] J. Cohen, "Trusses: Cohesive subgraphs for social network analysis," *National security agency technical report*, pp. 3–1, 2008.
- [11] A. Prat-Pérez, D. Dominguez-Sal, J. M. Brunat, and J.-L. Larriba-Pey, "Shaping communities out of triangles," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1677–1681.
- [12] J.-P. Eckmann and E. Moses, "Curvature of co-links uncovers hidden thematic layers in the world wide web," *Proceedings of the national academy of sciences*, vol. 99, no. 9, pp. 5825–5829, 2002.
- [13] N. Wang, J. Zhang, K.-L. Tan, and A. K. Tung, "On triangulation-based dense neighborhood graph discovery," *Proceedings of the VLDB Endowment*, pp. 58–68, 2010.
- [14] C. E. Tsourakakis, P. Drineas, E. Michelakis, I. Koutis, and C. Faloutsos, "Spectral counting of triangles via element-wise sparsification and triangle-based link recommendation," *Social Network Analysis and Mining*, pp. 75–81, 2011.
- [15] K. Date, K. Feng, R. Nagi, J. Xiong, N. S. Kim, and W.-M. Hwu, "Collaborative (CPU + GPU) algorithms for triangle counting and truss decomposition on the minsky architecture: Static graph challenge: Subgraph isomorphism," in *2017 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2017, pp. 1–7.
- [16] A. Yaşar, S. Rajamanickam, J. W. Berry, M. M. Wolf, J. Young, and Ü. V. Çatalyürek, "Linear algebra-based triangle counting via fine-grained tasking on heterogeneous environments," in *High Performance Extreme Computing Conference (HPEC), 2019 IEEE*. IEEE, 2019.
- [17] O. Green, P. Yalamanchili, and L.-M. Munguía, "Fast triangle counting on the GPU," in *Proceedings of the 4th Workshop on Irregular Applications: Architectures and Algorithms*. IEEE Press, 2014, pp. 1–8.
- [18] A. Yaşar, S. Rajamanickam, M. M. Wolf, J. W. Berry, and Ü. V. Çatalyürek, "Fast triangle counting using cilk," in *High Performance Extreme Computing Conference (HPEC), 2018 IEEE*, 2018, pp. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/document/8547563>
- [19] E. G. Boman, K. D. Devine, and S. Rajamanickam, "Scalable matrix computations on large scale-free graphs using 2d graph partitioning," in *SC'13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 1–12.
- [20] E. Saule, E. O. Bas, and Ü. V. Çatalyürek, "Load-balancing spatially located computations using rectangular partitions," *Journal of Parallel and Distributed Computing*, vol. 72, no. 10, pp. 1201–1214, 2012. [Online]. Available: <http://dx.doi.org/10.1016/j.jpdc.2012.05.013>
- [21] Ü. V. Çatalyürek, C. Aykanat, and B. Uçar, "On two-dimensional sparse matrix partitioning: Models, methods, and a recipe," *SIAM Journal on Scientific Computing (SISC)*, vol. 32, no. 2, pp. 656–683, 2010. [Online]. Available: <http://dx.doi.org/10.1137/080737770>
- [22] E. Kayaaslan, C. Aykanat, and B. Uçar, "1.5D parallel sparse matrix-vector multiply," *SIAM Journal on Scientific Computing*, vol. 40, no. 1, pp. C25–C46, 2018.
- [23] S. Acer, O. Selvitopi, and C. Aykanat, "Optimizing nonzero-based sparse matrix partitioning models via reducing latency," *Journal of Parallel and Distributed Computing*, vol. 122, pp. 145–158, 2018.
- [24] M. Grigni and F. Manne, "On the complexity of the generalized block distribution," in *International Workshop on Parallel Algorithms for Irregularly Structured Problems*. Springer, 1996, pp. 319–326.
- [25] A. Yaşar and Ü. V. Çatalyürek, "Heuristics for symmetric rectilinear matrix partitioning," ArXiv, Tech. Rep. arXiv:1909.12209, Oct 2019. [Online]. Available: <http://arxiv.org/abs/1909.12209>
- [26] D. P. O'leary and G. Stewart, "Data-flow algorithms for parallel matrix computation," *Communications of the ACM*, vol. 28, no. 8, pp. 840–853, 1985.
- [27] B. Hendrickson, R. Leland, and S. Plimpton, "An efficient parallel algorithm for matrix-vector multiplication," *International Journal of High Speed Computing*, vol. 7, no. 01, pp. 73–88, 1995.
- [28] E.-J. Im, K. Yelick, and R. Vuduc, "Sparsity: Optimization framework for sparse matrix kernels," *The International Journal of High Performance Computing Applications*, vol. 18, no. 1, pp. 135–158, 2004.

- [29] G. Teodoro, T. D. R. Hartley, Ü. V. Çatalyürek, and R. Ferreira, "Run-time optimizations for replicated dataflows on heterogeneous environments," in *Proc. of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC)*, 2010, pp. 13–24. [Online]. Available: <http://portal.acm.org/GraphBasedCitationAnalysis.cfm?id=1851479>
- [30] M. M. Wolf, M. Deveci, J. W. Berry, S. D. Hammond, and S. Rajamanickam, "Fast linear algebra-based triangle counting with kokkoskernels," in *IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2017, pp. 1–7.
- [31] M. Deveci, C. Trott, and S. Rajamanickam, "Performance-portable sparse matrix-matrix multiplication for many-core architectures," in *Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2017, pp. 693–702.
- [32] A. S. Tom and G. Karypis, "A 2D parallel triangle counting algorithm for distributed-memory architectures," in *International Conference on Parallel Processing*. ACM, 2019, p. 45.
- [33] S. Acer, A. Yaşar, S. Rajamanickam, M. M. Wolf, and Ü. V. Çatalyürek, "Scalable triangle counting on distributed-memory systems," in *High Performance Extreme Computing Conference (HPEC), 2019 IEEE*. IEEE, 2019.
- [34] Y. Hu, H. Liu, and H. H. Huang, "High-performance triangle counting on gpus," in *IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–5.
- [35] T. A. Davis and Y. Hu, "The University of Florida sparse matrix collection," *ACM Transactions on Mathematical Software (TOMS)*, p. 1, 2011.
- [36] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical programming*, vol. 91, no. 2, pp. 201–213, 2002.

Abdurrahman Yaşar is a PhD Student in the School of Computational Science and Engineering at Georgia Institute of Technology. He received his M.S. in Computer Engineering from Bilkent University, Turkey in 2015.

Sivasankaran Rajamanickam (M'14) is a Principal Member of Technical Staff in the Center for Computing Research at Sandia National Laboratories. He earned his B.E. from Madurai Kamaraj University, India, and his Ph.D. in Computer Engineering from University of Florida.

Jonathan Berry is a Distinguished Member of the Technical Staff at Sandia National Laboratories. He holds a Ph.D. in computer science from Rensselaer Polytechnic Institute and spent almost a decade in liberal arts academia before joining Sandia in 2004.

Ümit V. Çatalyürek (M'09,SM'10,Fellow'16) is a Professor and Associate Chair in the School of Computational Science and Engineering at the Georgia Institute of Technology. He received his Ph.D., M.S. and B.S. in Computer Engineering and Information Science from Bilkent University, Turkey.